New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

Collinearity, Heteroscedasticity and Outlier Diagnostics in Regression. Do They Always Offer What They Claim?

Germà Coenders and Marc Saez[1]

Abstract

In this article, some classic collinearity, heteroscedasticity and outlier diagnostics in multiple regression models are reviewed. Some major problems are described in the *Breusch-Pagan test*, the *condition number* and the critical values for the *studentized deleted residual* and *cook's distance*. Alternatives are suggested which consist of computing the condition number of the correlation matrix instead of the rescaled moment matrix, using the NR^2 statistic for the Breusch Pagan test, setting global-risk-based critical values for the studentized deleted residual, and drawing graphical displays for Cook's distance. Very large differences between the original and alternative diagnostics emerge both on simulated data and on real data from a work absenteeism study.

1 Introduction

This article will focus on three types of diagnostics for multiple regression models, namely collinearity, heteroscedasticity and outlier diagnostics. These diagnostics are probably the most crucial when analyzing cross-sectional data. For this type of data, dependence is less likely to occur and difficult to treat. Also non-normality is less critical as the number of observations (which is often limited in time series data) increases. Cross-sectional data often combine very small and very large units, which can be subject to different variability and which can inflate the correlations among all variables, including, of course, the regressors. Some of these large units may be outliers.

In this article, several collinearity, heteroscedasticity and outlier diagnostics of common use are first reviewed. We have detected some major problems in some

[1] Department of Economics, University of Girona. Faculty of Economics, Campus of Montilivi, 17071 Girona, Spain.

commonly used such diagnostics, namely the *Breusch-Pagan test* (Breusch and Pagan, 1979), the *condition number* (Belsley et al., 1980; Belsley, 1982), and the commonly used critical values for some outlier statistics such as Cook's distance (Cook, 1977) and the standardized deleted residual (Belsley et al., 1980). These problems are next described and alternatives are suggested. Finally, the classic diagnostics and the alternatives are compared on a real data set from a work absenteeism study (Saez et al., in press).

2 Classic collinearity, heteroscedasticity and outlier diagnostics

We consider a linear multiple regression model with *k* regressor variables:

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_k x_{ki}+u_i \qquad (1)$$

which can be expressed in matrix notation as:

$$\boldsymbol{y}=\boldsymbol{X}\beta+\boldsymbol{u} \qquad (2)$$

where $\boldsymbol{u} \rightarrow N(\boldsymbol{0},\sigma^2\boldsymbol{I})$

In multiple regression models, collinearity can be related to the existence of near linear dependencies among the columns of the **X** matrix. For each regressor $\mathbf{x}_j$, the *tolerance* can be computed as $Tol_j = 1\text{-}R^2_j$, where R^2_j is the coefficient of determination obtained in each of the *k* auxiliary regressions of the form:

$$x_{ji}=\delta_0+\delta_1 x_{1i}+...+\delta_{j-1}x_{j-1i}+\delta_{j+1}x_{j+1i}+...+\delta_k x_{ki}+v_i \qquad (3)$$

Thus, Tol_j shows the proportion of variance of $\mathbf{x}_j$ that is not accounted for by the remaining *k-1* regressors and can be used as an index of the degree of collinearity associated to $\mathbf{x}_j$.

Another index of collinearity of $\mathbf{x}_j$, called *variance inflation factor (Vif)*, can be obtained as a measure of the increment of the sampling variance of the estimated regression coefficient of $\mathbf{x_j}$ (b_j) due to collinearity. Vif_j can be computed as the *j*th diagonal value of the inverse of the **R** correlation matrix among the regressors or alternatively as $1/Tol_j$. Values of Vif_j lower than *10* or values of Tol_j larger than *0.1* are usually considered to be acceptable.

Overall measures of collinearity which take all regressors into account simultaneously have also been suggested. The most often used overall collinearity diagnostic is the *condition number* (Belsley et al., 1980; Belsley, 1982). The condition number of a matrix is the square root of the ratio of the largest to the

smallest eigen-values. A large condition number of the **X'X** *augmented moment matrix*, reflects the existence of one or more near linear dependencies among the columns of **X** (Belsley et al., 1980).

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \qquad (4)$$

In order to avoid the dependence of eigen values on the scaling of the data it is common practice to first normalize the **X'X** matrix by pre and post-multiplying by a diagonal matrix containing the square root of the moments about zero of all regressors including the constant term (Greene, 1993). A *rescaled augmented moment matrix* can be computed as **SX'XS**, where **S** is a diagonal matrix whose *j*th element is $1/(\mathbf{x'_j x_j})$ where $\mathbf{x_j}$ is a column vector with the values of the *j*th regressor. Values of the condition number of **SX'XS** lower than *30* are usually considered to be acceptable.

The most often used heteroscedasticity diagnostics are statistical tests of the null homoscedasticity hypothesis against the alternative that a function of the variance of the *i*th disturbance σ^2_i can be linearly or non-linearly related to a set of **z** variables. Among them, the most often used are the Breusch-Pagan test (Breusch and Pagan, 1979) and the *White test* (White, 1980). Some others were suggested by Harvey (1976) and Glesjer (1969).

The alternative hypothesis in the Breusch-Pagan test is that the variance of the disturbance is linearly related to the set of **z** variables, which may or may not coincide with the set of **x** regressor variables:

$$\sigma^2_{ui} = \eta_0 + \eta_1 z_{1i} + \eta_2 z_{2i} + \ldots + \eta_m z_{mi} \qquad (5)$$

The test assesses the joint significance of the $\eta_1,\ldots,\eta_m$ parameters by first estimating the model in Equation 1 by ordinary least squares, then squaring the standardized residuals (divided by the maximum likelihood estimate of σ obtained assuming that **u** is normal and homoscedastic). Next these squared standardized residuals are regressed on the $\mathbf{z_1},\ldots,\mathbf{z_m}$ variables, what we call auxiliary regression.

It can be shown that, if $\eta_1=\eta_2=\ldots=\eta_m=0$, then NR^2 is asymptotically distributed as a chi-square with *m* degrees of freedom, where *N* is the number of observations and R^2 is the coefficient of determination of the auxiliary regression. If **u** is normal and homoscedastic, the asymptotic variance of the squared standardized residuals is equal to *2* and *SSQR/2* is also distributed as a chi-square with *m* degrees of freedom, where *SSQR* is the explained sum of squares in the auxiliary regression. *SSQR/2* is the test statistic suggested by Breusch and Pagan in 1979 and is equivalent to a Lagrangian Multiplier test statistic under normality. The use of the NR^2 statistic was suggested by Koenkar (1981) and Evans (1992).

The White test is analogous to the Breusch-Pagan test except for two differences. First, the set of **z** variables is constituted by the squared **x** variables and all possible second-order interactions among them. Second, NR^2 is the suggested test statistic.

From a practical perspective, two types of outliers are problematic in regression analysis. On the one hand, some observations may fail to be predicted by the model with a reasonable degree of accuracy. This type of outliers may reveal the fact that several populations are mixed in the data set or that some relevant variables have been omitted. On the other hand, some observations may be influential in the sense that their presence in the data set substantially modifies the estimates. This type of outliers weakens the conclusions which may be drawn from the model. Of course it often happens that an observation is an outlier according to both definitions simultaneously. Robust regression procedures (Chen and Dixon, 1972) may be used as a safeguard against outliers without needing to test for their presence, but the potential information given by outliers about mixed populations or omitted variables will be missed.

A hard to predict observation which is not influential may be detected from its large residual. However, a hard to predict observation which is influential will often have a small residual. The *studentized deleted residual* is then suggested as an alternative by Belsley et al. (1980). The studentized deleted residual of the *i*th observation is the residual computed from a regression equation estimated without the *i*th observation divided by its standard deviation, which is also computed without the *i*th observation. This prevents the *i*th observation from influencing its own prediction and from inflating the standard error with which it is being standardized. The studentized deleted residual can be computed as:

$$\frac{y_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{1+h_{(ii)}}} \qquad (6)$$

where $\hat{y}_{(i)}$ is the predicted value from the estimates of the regression omitting the *i*th observation, $s_{(i)}$ is the least squares estimate of σ obtained in a regression omitting the *i*th observation and $h_{(ii)}$ is $\mathbf{x_i(X_{(i)}'X_{(i)})^{-1}x_i'}$, where $\mathbf{x_i}$ is the row vector of regressors for the *i*th observation and $\mathbf{X_{(i)}}$ the matrix of regressors of all observations except the *i*th.

Cook's distance (Cook, 1977) is the usual statistic which is employed to detect influential observations. The Cook's distance associated to the *i*th observation is a standardized distance measure between the vectors of regression slope estimates obtained with and without the *i*th observation. It can also be computed as a function of the residual and the so-called *leverage value*.

$$\frac{1}{k+1}\frac{h_{ii}}{(1-h_{ii})^2}\frac{e_i^2}{s^2} \qquad (7)$$

where h_{ii} is the leverage value computed as $\mathbf{x_i(X'X)^{-1}x_i'}$, e_i is the residual and s is the least squares estimate of σ.

In this article we will concentrate on studentized deleted residuals and Cook's distances. In order to reduce the arbitrariness of the interpretation of these statistics, the use of some sort of critical values has been suggested. If the normality assumption holds, then studentized deleted residuals follow a Student's t distribution, so that for reasonable sample sizes critical values may be selected according to the desired percentile of the standardized normal distribution. Critical levels of ±1.96 or ±2 (5% risk) and ±3 (0.27% risk, typical in process control charts) have been suggested (Belsley et al., 1980; Greene, 1993).

Critical levels for Cook's distance are usually based on non-probabilistic criteria. In order to use probabilistic critical values, a particular multivariate distribution model should be assumed for **X**, which would often be unreasonable. Cook (1977) and Weisberg (1980) suggest using the *50*th percentile of the *F* distribution with *k* and *N-k-1* degrees of freedom.

In the next section we report some major problems in the condition number, the Breusch-Pagan test and the usual critical values for Cook's distances and the studentized deleted residuals. Alternatives will be suggested.

3 Critique of some classic diagnostics

3.1 Critique of the condition number

In this subsection we show that the condition number of **SX'XS** is heavily dependent on the means of the regressor variables. If the means of some or all of the regressors are high in comparison to their standard deviations, the condition number will convey an apparently large degree of collinearity, even if the regressors are in fact orthogonal. This is due to the fact that the off-diagonal elements of the augmented moment matrix can be very large as they contain the product of the means of the variables involved.

We will illustrate this problem with an artificial data set which contains one dependent variable **y** and two orthogonal regressors $\mathbf{x_1}$ and $\mathbf{x_2}$. The data set has not been generated to follow any specific distribution but only in order to ensure exact orthogonality and rounded up computations.

The regressors may be considered to be measured as index numbers (with base *100*). They are once considered in their original form (thus having a mean value somewhere around *100*, see Table 1) and once after the constant value *100* has been subtracted (thus having a mean value somewhere around *0*, see Table 2). The large differences in the **SX'XS** matrices makes the condition number yield

completely different values on the same data depending on whether the constant *100* is subtracted from all regressors or not.

Table 1: Original data set, with correlation matrix and rescaled moment matrix.

Raw Data

X1	X2	Y
101	101	9
101	102	10
101	103	11
102	101	9
102	102	11
102	103	13
103	101	9
103	102	12
103	103	15

Correlation Matrix (**R**)

	X1	X2
X1	1.0000	0.0000
X2	0.0000	1.0000

Rescaled Augmented Moment Matrix (**SX'XS**)

	Const	X1	X2
Const	1.000000	0.999968	0.999968
X1	0.999968	1.000000	0.999936
X2	0.999968	0.999936	1.000000

Table 2: Data set once 100 has been subtracted from both regressors, with correlation matrix and rescaled moment matrix.

Raw Data

X1	X2	Y
1	1	9
1	2	10
1	3	11
2	1	9
2	2	11
2	3	13
3	1	9
3	2	12
3	3	15

Correlation Matrix (**R**)

	X1	X2
X1	1.0000	0.0000
X2	0.0000	1.0000

Rescaled Augmented Moment Matrix (**SX'XS**)

	Const	X1	X2
Const	1.00000	0.92582	0.92582
X1	0.92582	1.00000	1.85714
X2	0.92582	0.85714	1.00000

The results of the *ordinary least squares* (OLS) estimation on the data in Table 1 are displayed in Table 3 and convey a very good fit and significance of the variables. Note the huge value of the condition number (most authors suggest that values above 30 show an extremely high degree of collinearity). On the contrary, the *Vif*'s correctly reflect the exact orthogonality of the regressors.

Table 4 shows the same results obtained from the data in Table 2. Once *100* has been subtracted from both regressors, the condition number, while not equal to

1, takes a value which most researchers would consider to indicate an acceptable level of collinearity. The estimates, standard errors and *t*-values of the regressors are identical and equally good to those in Table 3; only the results for the constant term change.

Table 3: OLS estimates and diagnostics on the data in Table 1.

General Diagnostics		Estimates and Variable-Specific			
Diagnostics					
Explained sum of squares	30	Variable	b	Vif	t
Residual sum of squares	4	X1	1.000	1.000	3.000
R^2	0.882	X2	2.000	1.000	6.000
Condition number of **SX'XS**	374.773	Const	-295.000	-6.135	

Table 4: OLS estimates and diagnostics on the data in Table 2.

General Diagnostics		Estimates and Variable-Specific			
Diagnostics					
Explained sum of squares	30	Variable	b	Vif	t
Residual sum of squares	4	X1	1.000	1.000	3.000
R^2	0.882	X2	2.000	1.000	6.000
Condition number of **SX'XS**	7.425	Const	5.000		5.095

We then discourage the use of the condition number whenever some of the regressors have a high mean. As an alternative one can still use other usual collinearity diagnostics as the *Vif*, and the *Tol*. One drawback of these diagnostics is the fact that they only evaluate regressors one by one. If one is willing to use one single global measure of collinearity we suggest to compute the condition number of the **R** correlation matrix among the regressors instead of **SX'XS**. In our example, the correlation matrix is a *2×2* identity matrix. Both eigen-values are equal to *1* and the suggested condition number is also equal to *1*, thus indicating a total absence of collinearity. This procedure is easier than, for instance, the generalised collinearity diagnostics suggested by Fox and Monette (1992) and the signal-to-noise test (Belsley, 1982).

3.2 Critique of the Breusch-Pagan test

In this subsection we review the strong dependence of the Breusch-Pagan test on the normality assumption of the disturbances (Evans, 1992). If **u** is homoscedastic but non-normal, the variance of the squared standardized residuals will be different from *2* (larger under leptokurtosis and lower under platykurtosis) thus distorting the test results when using the *SSQR/2* statistic (Greene, 1993). In particular, the Breusch-Pagan test often leads to the rejection of the true homoscedasticity hypothesis when **u** is leptokurtic.

To illustrate the sensitivity of the Breusch-Pagan test on the kurtosis of **u**, we carried out a limited Monte Carlo experiment. We simulated a simple regression model and carried out the standard Breusch-Pagan test and a modified test using NR^2 including the same $\mathbf{x_1}$ regressor in the main and auxiliary regressions. $\mathbf{x_1}$ had a discrete uniform distribution, taking 5 consecutive integer values. The R^2 of the main regression was 80% and the number of replications 500. The simulation was carried out under a number of different conditions:

1. The disturbances of the main regression could be homoscedastic, moderately heteroscedastic (variance proportional to $\mathbf{x_1}$ where $\mathbf{x_1}$ ranged from 11 to 15) or strongly heteroscedastic (variance proportional to $\mathbf{x_1}$ where $\mathbf{x_1}$ ranged from 4 to 8).
2. The distribution of **u** could be platykurtic (uniform), mesokurtic (normal), moderately leptokurtic (Student's *t* with 5 d.f.) or strongly leptokurtic (Student's *t* with 3 d.f.).
3. The number of observations could be *N=500* or *N=100*.

The dependent variables in the experiment are the percentage of times the *SSQR/2* and NR^2 statistics exceed the 5% critical value of the chi-square distribution with one degree of freedom. The results are displayed in Table 5. The first column of Table 5 contains the rejection rates when the null homoscedasticity hypothesis is true. There we see that the NR^2 statistic always yields a rate which is close to the theoretical 5%, even when **u** is not normal or the small number of observations might make the asymptotic approximation suspect. On the contrary, the original Breusch-Pagan test only approaches the theoretical rejection rate if **u** is normal. The rejection rate approaches 0 when **u** is platykurtic, and gets very large when **u** is leptokurtic, specially when kurtosis is high. The Breusch-Pagan test is then very sensitive to leptokurtic disturbances. As a matter of fact, the test could even be used to test for kurtosis, even if there may be better alternatives available (Mardia, 1974).

The remainder of Table 5 represents the case where the alternative hypothesis holds, under various degrees of power arising from different sample sizes and degrees of heteroscedasticity. The power is larger for NR^2 when **u** is platykurtic

and larger for *SSQR/2* when *u* is leptokurtic, although the latter is in some sense fallacious, as both hypotheses against which the test is sensitive are simultaneously false, so that not all of the rejection rate can be attributed to power against heteroscedasticity. The fact that the power of both statistics seems to be the about the same when **u** is normal, is a further argument for using the NR^2 statistic.

Table 5: Results of the Monte Carlo experiment. Percentage of replications in which the *SSQR/2* and NR^2 statistics exceed *3.84.*

		Homoscedasticity		Moderate Heteroscedasticity		Strong Heteroscedasticity	
N	**u** Distr.	NR^2	*SSQR/2*	NR^2	*SSQR/2*	NR^2	*SSQR/2*
500	uniform	.042	.002[a]	.794	.360	1	.992
500	normal	.034	.032	.388	.388	.950	.946
500	*t* with 5 d.f.	.042	.200[a]	.224	.468	.678	.894
500	*t* with 3 d.f.	.044	.470[a]	.104	.526	.296	.738
100	uniform	.072[a]	.006[a]	.208	.028	.668	.264
100	normal	.046	.040	.100	.086	.330	.326
100	*t* with 5 d.f.	.046	.162[a]	.082	.236	.220	.430
100	*t* with 3 d.f.	.038	.314[a]	.058	.326	.138	.432

[a] Rates significantly different (α=5%, two tailed) from the theoretical .05 value in the homoscedasticity case.

The original Breusch-Pagan test will then be particularly misleading if **u** is leptokurtic and homoscedastic. If a considerable number of replicates are available for some combinations of values of the regressors, normality can be tested prior to using the original Breusch-Pagan test. Otherwise, we suggest that the alternative NR^2 statistic should generally be used, thus bringing the Breusch-Pagan test closer to White's without suffering the problems of White's test when *k* is large (e.g. if *k = 10*, White's procedure implies that *55* regressors be included in the auxiliary regression, which can lead to a substantial reduction in power). The use of the NR^2 statistic does not assume zero kurtosis but just that kurtosis is constant for all observations (White, 1980).

3.3 Critique of critical values for outlier statistics

The main critique about probabilistic limits, such as those for studentized deleted residuals is that they are usually based on individual risk and not on overall risk. In other words, the distribution of an individual observation cannot be used to assess the characteristics of extreme values (Barnett and Lewis, 1994). For instance, if we use the ± 1.96 limits, then the individual risk of $\alpha_i=5\%$ but the global risk can be much larger for large N. For instance, if $N=1000$ we will expect around 50 residuals to be larger than the critical value even if no outliers at all are present. The risk that there will be at least one residual above the critical value under these circumstances is nearly $\alpha_g=100\%$.

Under the normality assumption, the squared studentized deleted residuals are asymptotically distributed as a chi-square with one degree of freedom and asymptotically independent. In this case, the global risk can be computed as

$$\alpha_g=1-(\phi(l^2))^N \qquad (8)$$

where ϕ is the distribution function of a chi-square with one degree of freedom and l is the critical value for the studentized deleted residual. We suggest always using critical values that yield a reasonable global risk α_g selected by the user. Such limits can be computed as:

$$\ell = \sqrt{\phi^{-1}(\sqrt[N]{1-\alpha_g})} \qquad (9)$$

where ϕ^{-1} is the inverse of the distribution function of a chi-square with one degree of freedom. Tests based on studentized deleted residuals have been proven to be likelihood ratio tests for normally distributed disturbances (Barnett and Lewis, 1994). For small samples, independence of the residuals of different observations is not warranted and critical values based on Bonferroni's inequality are available (Lund, 1975).

As regards Cook's distance, we suggest not using any limits at all but drawing a scree plot of Cook's distances ordered from highest to lowest. This plot will be useful to separate the few most influential observations from the many least influential ones. A sensitivity analysis of the estimates should then be carried out by hand by sequentially dropping the identified observations and qualitatively evaluating the extent to which the conclusions to be drawn from the model change. This is what ultimately counts when evaluating influence of the observations and is far more useful than blindly using a fixed critical value. Alternatively, for each parameter's conficence interval one could report the lowest value for the lower limits and the highest value for the upper limits found when dropping different observations (Leamer, 1979).

4 Illustration of the classic and alternative diagnostics

In this section we illustrate the use of all reviewed collinearity and heteroscedasticity diagnostics on real data and compare them to those suggested as an alternative to the classic ones. The illustration is carried out on a data set from a work absenteeism study (Saez et al., in press). The aims of the study were to determine the factors that explain work absences due to illness or accident at the working place in the public transport company in Barcelona, Spain between 1994 and 1996. The data were measured for the whole population of over *5,000* employees. The data presented in this article do not exactly coincide with those used in Saez et al. In order to bring statistical power within the boundary of usual standards, we drew a random sample of *500* employees, which were reduced to *431* after list-wise deletion of missing cases. For the sake of simplicity we restricted ourselves to a subset of the variables (the ones which proved to have the most theoretical and statistical significance in the original study) and we grouped the categories of some qualitative regressor variables. This grouping was made by collapsing categories which were similar from a theoretical point of view and which had a similar effect on the dependent variable in the original study.

The dependent variable is labelled *spells* and measures the number of work absences of the employee during the period ranging from 1994 to 1996. Even if it is a discrete count variable, it varies over a range which is reasonably wide in order to be used as dependent variable in a standard regression model (see histogram in Figure 1).

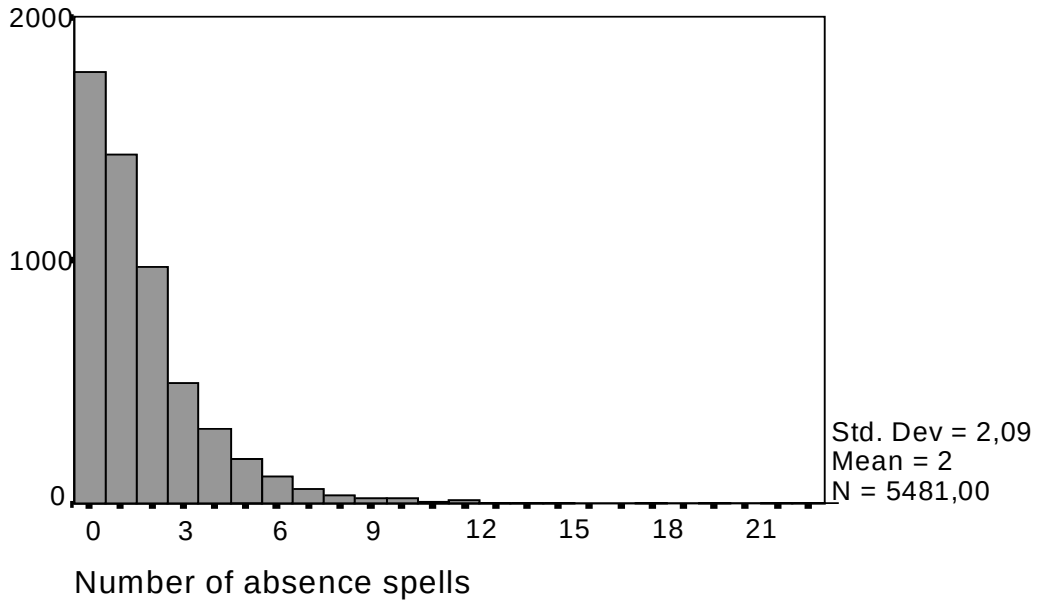

Figure 1: Histogram of the dependent variable for the whole population.

The regressor variables were:

1. *Age* (in years).
2. *Tenure* (in years).
3. *Ocup:* dummy (1: electricians, drivers and mechanic; 0: others).
4. *Shift:* dummy (1:day; 0: night or mixed).
5. *Study:* dummy (1:primary or less; 0: secondary or higher).
6. *Tobacco:* dummy (1:smokers and former smokers; 0: never smoked).
7. *Disease:* dummy (1:one or more serious illnesses; 0: none).
8. *Gender:* dummy (1:female; 0 male).
9. *Selfper:* dummy (1:bad perceived health; 0: good perceived health).
10. *Company:* dummy (1:Underground; 0 Bus).

Table 6: Distribution of the regressor variables.

Variable	Cases	Mean	Std Dev
Age	470	43.65	11.31
Tenure	470	16.85	10.82
Ocup	470	.51	.50
Shift	470	.44	.50
Study	470	.65	.48
Tobacco	470	.58	.49
Disease	470	.44	.50
Gender	470	.10	.30
Selfper	431	.01	.12
Company	470	.45	.50

Correlation Coefficients

	Age	Tenur	Ocup	Shift	Study	Tobac	Disea	Gend	Selfp	Comp
Age	1.00	.88	-.14	.23	.66	.02	.20	-.10	.08	.13
Tenur	.88	1.00	-.20	.32	.58	.00	.23	-.07	.08	.12
Ocup	-.14	-.20	1.00	-.51	-.06	-.04	.08	-.33	-.01	-.63
Shift	.23	.32	-.51	1.00	.16	-.01	.03	.16	.10	.50
Study	.66	.58	-.06	.16	1.00	.02	.16	-.12	.05	.07
Tobac	.02	.00	-.04	-.01	.02	1.00	-.12	-.06	.03	.07
Disea	.20	.23	.08	.03	.16	-.12	1.00	-.13	.05	-.11
Gend	-.10	-.07	-.33	.16	-.12	-.06	-.13	1.00	-.04	.22
Selfp	.08	.08	-.01	.10	.05	.03	.05	-.04	1.00	.09
Comp	.13	.12	-.63	.50	.07	.07	-.11	.22	.09	1.00

The distribution of the regressor variables is shown in Table 6. Some correlations are substantial but none is extremely high (the highest is *.88*), which suggests that collinearity is not extreme. Note that the mean of the variable *age* is relatively high compared to the standard deviation, which is likely to inflate the condition number of **SX'XS** with respect to that of **R**.

Table 7: Estimates and diagnostics of the regression model (n=431).

Variable	b	Se(b)	Tol	Vif	t
Age	.004	.024	.159	6.264	.171
Tenure	.009	.025	.168	5.937	.382
Ocup	1.005	.310	.491	2.037	3.238
Shift	.368	.281	.607	1.647	1.310
Study	.120	.302	.550	1.817	.398
Tobacco	.288	.221	.974	1.027	1.302
Disease	.351	.232	.878	1.139	1.510
Gender	1.161	.408	.840	1.190	2.843
Selfper	2.847	.939	.973	1.027	3.031
Company	1.205	.306	.511	1.959	3.941
Const	-.282	.735			-.384

Diagnostics:

R^2	.114
Condition number of **SX'XS**:	30.480
Condition number of **R**:	5.799
Kurtosis of residuals:	31.033
SSQR/2 Breusch-Pagan test statistic:	164.523 (p-value: 0.00000)
NR^2 Breusch-Pagan test statistic:	10.611 (p-value: 0.38956)

Studentized deleted residuals larger than the critical value:

Critical value	Count	α_i	α_g
±1.96	15	5%	100%
±3	3	0.27%	69%
±3.66	2	0.025%	10%
±3.85	2	0.012%	5%

The ordinary least squares estimates are in Table 7, together with the collinearity and heteroscedasticity diagnostics discussed in previous sections. The *Vif*'s and the *Tol*'s show a sizeable though not extreme degree of collinearity. However, the condition number of **SX'XS** lies above the usually recommended

limits, thus showing an apparently extreme collinearity. The condition number of **R**, on the contrary, shows a large though not extreme degree of collinearity, in coherence with the *Vif*'s and the *Tol*'s. The extremely high kurtosis of the residuals causes big difference between the *SSQR/2* and NR^2 statistics for the Breusch-Pagan test. There is no grounds to reject the null homoscedasticity hypothesis when using the alternative NR^2 statistic. On the contrary, the *SSQR/2* statistic would lead to the rejection of the null hypothesis with virtually no risk, although it is unknown whether the test is being sensitive to heteroscedasticity or to kurtosis.

Table 7 also shows the number of observations whose studentized deleted residual exceeds a range of alternative critical values. The number of outliers detected changes dramatically from *15* to *2* depending on whether α_i or α_g is set to *5%*. Even the *±3* critical values have an unacceptably large α_g. A global risk α_g*=10%* seems a reasonable compromise if we want individual outliers to have a greater chance of being detected.

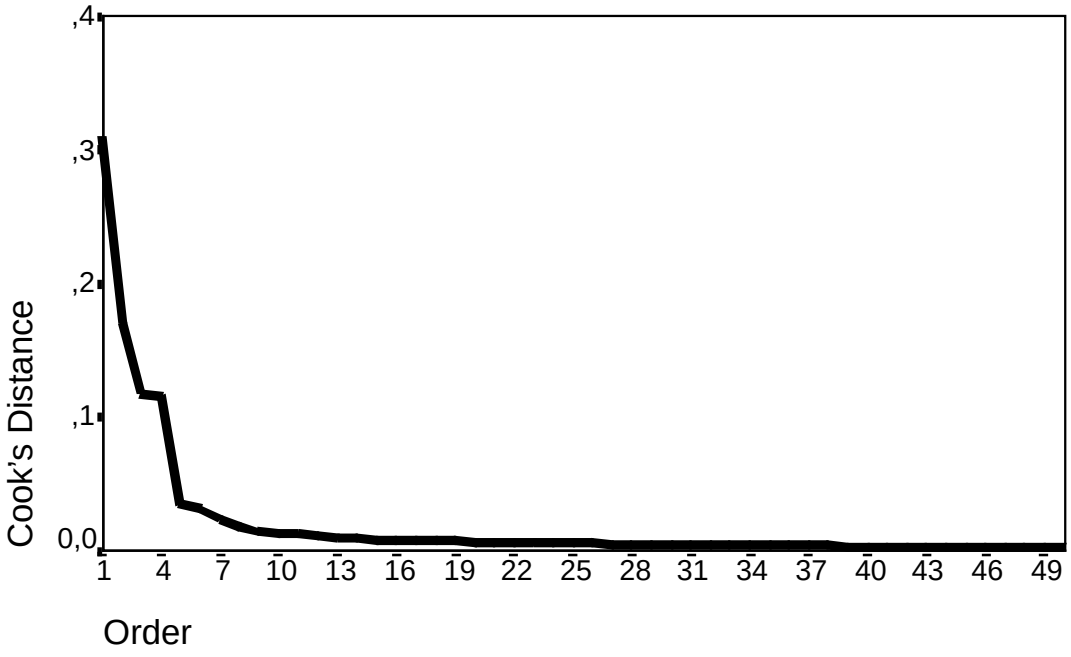

Figure 2: Scree plot of Cook's distances (only the 50 largest are shown).

Figure 2 shows a scree plot of the *50* largest Cook's distances. The plot suggests that *4* observations are comparatively more influential than most others and could be subject to a sensitivity analysis. The standard level based on the *50*th percentile of the *F* distribution with *11* and *420* degrees of freedom is *0.94*, which is not exceeded by any of the observations. The removal of the *4* observations with

the highest Cook's distances did lead to some substantial changes in the model: the variable *tobacco* attained a significance of *5.5%*; the R^2 statistic increased to *.142*; the estimated effect of *gender* decreased to *0.75* and the effect of *selfper* increased to *3.22*.

5 Discussion

In this article we have shown major weaknesses of some commonly used collinearity heteroscedasticity and outlier diagnostics. In particular, the condition number of **SX'XS** has been shown to be sensitive to the means of the variables, the Breusch-Pagan test has been shown to be sensitive to the kurtosis of the disturbances and the critical values for the usual outlier statistics have been shown to be rather meaningless. Alternatives which do not have these undesirable properties and which are easy to compute have been suggested.

The classic diagnostics have been compared to the alternatives on an empirical data set. The differences were large enough to lead to completely different conclusions depending on which diagnostics were employed.

Further robustness problems of these diagnostics are not solved by the suggested alternatives. Critical values for studentized deleted residuals are very sensitive to the normality assumption. Heteroscedasticity tests are very sensitive to the presence of outliers because they involve squaring the residuals, which makes outliers to have a more serious effect in the auxiliary than in the main regression.

References

[1] Barnett, V. and Lewis, T. (1994): *Outliers in Statistical Data.* New York: John Wiley & Sons.

[2] Belsley, D.A. (1982): Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics,* **20**, 211-253.

[3] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980): *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity.* New York: John Wiley & Sons.

[4] Breusch, T.S. and Pagan, A.R. (1979): A simple test for heteroscedasticity and random coefficient variation. *Econometrica,* **47**, 1287-1294.

[5] Chen, E.H. and Dixon, W.J. (1972). Estimates of parameters of a censored regression sample. *Journal of the American Statistical Association,* **6,** 664-671.

[6] Cook, R.D.(1977): Detection of influential observations in linear regression. *Technometrics,* **19**, 15-18.

[7] Evans, M. (1992): Robustness of size of tests of autocorrelation and heteroscedasticity to nonnormality. *Journal of Econometrics,* 7-24.

[8] Fox, J. and Monette, G. (1992): Generalized collinearity diagnostics. *Journal of the American Statistical Association,* **87**, 178-183.

[9] Glesjer, H. (1969): A new test for heteroscedasticity. *Journal of the American Statistical Association,* **64,** 316-323.

[10] Greene, W.H. (1993): *Econometric Analysis.* Englewood Cliffs, NJ: Prentice-Hall.

[11] Harvey, A.C. (1976): Estimating regression models with multiplicative heteroscedasticity. *Econometrica,* **44**, 461-465.

[12] Koenkar, R. (1981): A note on studentizing a test for heteroscedasticity. *Journal of Econometrics,* **17**, 107-112.

[13] Leamer, E. (1979): *Specification Searches in Econometrics.* New York: John Wiley & Sons.

[14] Lund, R.E. (1975): Tables for an approximate test for outliers in linear models. *Technometrics*, **17**, 473-476.

[15] Mardia, K.V. (1974): Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya B,* **36**, 115-128.

[16] Saez, M., Barceló, M.A., and Benavides, S. G. (in press): Análisis longitudinal de la incidencia de las ausencias de trabajo entre trabajadores de transporte urbano [Longitudinal analysis of the incidence of work absenteeism among public transport workers]. *Revista de Economía Aplicada.*

[17] Weisberg, S. (1980): *Applied Linear Regression.* New York: John Wiley & Sons.

[18] White, H. (1980): A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,* **48**, 817-838.